

Can a CNN Recognize Catalan Diet?

Pedro Herruzo^{a)}, Marc Bolaños^{b)} and Petia Radeva^{c)}

*Universitat de Barcelona. Barcelona, Spain.
Computer Vision Center. Bellaterra, Spain.*

^{a)}pherrusa7@alumnes.ub.edu

^{b)}marc.bolanos@ub.edu

^{c)}petia.ivanova@ub.edu

Abstract. Nowadays, we can find several diseases related to the unhealthy diet habits of the population, such as diabetes, obesity, anemia, bulimia and anorexia. In many cases, these diseases are related to the food consumption of people. Mediterranean diet is scientifically known as a healthy diet that helps to prevent many metabolic diseases. In particular, our work focuses on the recognition of Mediterranean food and dishes. The development of this methodology would allow to analyse the daily habits of users with wearable cameras, within the topic of lifelogging. By using automatic mechanisms we could build an objective tool for the analysis of the patient's behaviour, allowing specialists to discover unhealthy food patterns and understand the user's lifestyle.

With the aim to automatically recognize a complete diet, we introduce a challenging multi-labeled dataset related to Mediterranean diet called FoodCAT. The first type of label provided consists of 115 food classes with an average of 400 images per dish, and the second one consists of 12 food categories with an average of 3800 pictures per class. This dataset will serve as a basis for the development of automatic diet recognition. In this context, deep learning and more specifically, Convolutional Neural Networks (CNNs), currently are state-of-the-art methods for automatic food recognition. In our work, we compare several architectures for image classification, with the purpose of diet recognition. Applying the best model for recognising food categories, we achieve a top-1 accuracy of 72.29%, and top-5 of 97.07%. In a complete diet recognition of dishes from Mediterranean diet, enlarged with the Food-101 dataset for international dishes recognition, we achieve a top-1 accuracy of 68.07%, and top-5 of 89.53%, for a total of 115+101 food classes.

INTRODUCTION

Technology that helps track health and fitness is on the rise, in particular, automatic food recognition is a hot topic for both, research and industry. People around us have at least 2 devices, such as tablets, computers, or phones, which are used daily to take pictures. These pictures are commonly related to food; people upload dishes to social networks such as Instagram, Facebook, Foodspotting or Twitter. They do it for several reasons, to share a dinner with a friend, to keep track of a healthy diet or to show their own recipes. This amount of pictures is really attractive for companies, who are already putting much effort to understand people's diet, in order to offer personal food assistance and get benefits.

Food and nutrition are directly related to health. Obesity, diabetes, anemia, and other diseases, are all closely related to food consumption. Looking at food habits, the Mediterranean diet is scientifically known as a healthy diet. For example, a growing number of scientific researches has been demonstrating that olive oil, operates a crucial role on the prevention of cardiovascular and tumoral diseases, being related with low mortality and morbidity in populations that tend to follow a Mediterranean diet [1]. Many doctors tell patients to write a diary of their diet, trying to make them aware of what they are eating. Usually people do not care too much about that, annotating all the meals often is getting boring. An alternative is to make the food diary by pictures with the phone, or even better, to take the pictures automatically with a small wearable camera. It can be very useful in order to analyse the daily habits of users with wearable cameras. It appears as an objective tool for the analysis of patient's behaviour, allowing specialists to discover unhealthy food patterns and understand user's lifestyle. However, automatic food recognition and analysis are still challenges to solve for the computer vision community.

Deep learning and more specifically, Convolutional Neural Networks (CNNs) are actually the technologies within



FIGURE 1. Examples of Catalan cuisine in *FoodCAT* dataset: sauteed beans, paella, strawberries with vinegar, cuttlefish with peas, roasted snails and beans with sausage.

the state-of-the-art for automatic food recognition. The *GoogLeNet* [2] was responsible for setting the state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge in 2014 *ILSVRC14* [3]. Another widely used model is *VGG* [4], which secured the first and the second places also for the ImageNet *ILSVRC14* competition [3], in the localization and classification tasks respectively. One of the most popular food dataset is the *Food-101* dataset [5], containing 101 food categories, with 101.000 images. Another well known is the *UEC FOOD 256* dataset [6], which contains 256 types of food. Many researchers have been working with these datasets achieving very good results on food recognition [7], or in both food localisation and recognition [8] [9]. Another food related classification task that we are interested in, is to classify food categories, e.g. we should be able to classify a paella picture into the category of rice. In our case, we will do it following a robust classification of Catalan diet proposed in the book *El Corpus del patrimoni culinari català* [10]. Other related works on that topic classify 85 food classes [11] or 50 dishes [12]. Hence, we construct our dataset from the Catalan cuisine as a good representative of the Mediterranean food.

In this paper we focus on developing automatic algorithms to recognize Catalan food using deep learning techniques. For this purpose we build a dataset and enlarge it with the public domain dataset *Food-101*. Our work is organized in three steps:

1. *Build a dataset including healthy food:* The current food datasets are built in order to achieve a good performance in the general challenge of recognizing pictures automatically. Our goal is to present a method for food recognition of extended dataset based on Catalan food, as it is scientifically supported as a healthy diet (see Fig. 1 for some examples). Therefore, we present a new dataset based on Catalan food, which we call *FoodCAT*. This dataset has been classified following two different approaches. On one side, the images have been classified based on dishes, and on the other side, in a more general food categories. As an example, our system will recognize a dish with chickpeas with spinach as the food class 'chickpeas with spinach', but also as food category 'Legumes'.

2. *Recognize food dishes with Convolutional Neural Networks:* We are interested in applying a Convolutional Neural Network to recognize the new built healthy dataset together with the dataset *Food-101* [5]. We use pre-trained models over the large dataset *ImageNet*, such as *GoogLeNet* [2] and the *VGG* [4]. Moreover, in order to recognize food categories, we compare the differences between fine-tuning a pre-trained model over all the layers, versus the same model trained only for the last fully-connected layer.

3. *Improve the quality of the dataset and the recognition task with Super-Resolution:* It has been proven that large image resolution improves recognition accuracy [13]. Therefore, we will base on a new method to increase the resolution of the images, based on a Convolutional Neural Network, known as Super-Resolution (SR) [14]. With that, our goal is to get a better performance in the image recognition task.

METHODOLOGY

The image classification problem is the task of assigning a label from a predefined set of categories to an input image. In order to tackle this task for the Catalan diet problem, we propose taking a data-driven approach. After collecting a dataset for the problem at hand, we are going to train a CNN for automatically learning the appearance of each class and classifying them.

The collected dataset, named *FoodCAT*, when compared to the most widely used dataset for food classification *Food-101*, presents a lower image resolution which, as we prove in our experiments, leads to a data bias and a lower performance when training a CNN on the combined datasets. In order to solve this problem, we must increase the resolution to at least 256x256 pixels, which is the usual input size to CNNs. Thus, we propose using the method known as *Super-Resolution* and consequently improve the accuracy in the food recognition task.

Model

In order to apply food classification, we propose using the *GoogleNet* architecture, which has proven to obtain very high performance in several classification tasks [7] [8] [15].

We train the *GoogleNet* model using an image crop of 224x224x3 pixels as input. During training, in order to perform data augmentation, we extract random crops from the images after unifying their resolution to 256x256x3. During the testing procedure, we use the central image crop. The *GoogleNet* convolutional neural network architecture is a replication of the model described in the *GoogleNet* publication [2]. The network is 22 layers deep when counting only layers with parameters (or 27 layers if we also count pooling layers). As the authors explain in their paper [2], two of the features that made this net so powerful are : *Auxiliary classifiers connected to the intermediate layers*: which was thought to combat the vanishing gradient problem given the relatively large depth of the network. During training, their loss gets added to the total loss of the network with a discount weight. In practice, the auxiliary networks effect is relatively minor (around 0.5%) and it is required only one of them to achieve the same effect. *Inception modules*: the main idea for it is that in images, correlations tend to be local. Therefore, in each of the 9 modules, they use convolutions of dimension 1x1, 3x3, 5x5, and pooling layers of 3x3. Then, they put all outputs together as a concatenation. Note that to reduce the depth of the volume, convolutions 3x3 and 5x5 are performed after applying a 1x1 convolution with less filters, and pooling 3x3 is also followed by a convolution 1x1. This makes the model more efficient reducing the number of parameters in the net.

Super-Resolution

The image dimensions of *FoodCAT* dataset are on average smaller than 256x256. Motivated by the fact that larger images improve recognition accuracy [13], we propose increasing the resolution with a state-of-the-art method instead of applying a common upsampling through bilinear interpolation. To increase the size of the images, we use the method called Super-Resolution [14]. In this paper, the authors propose a technique for obtaining a High-Resolution (HR) image from a Low-Resolution (LR) one. To this end, they use a Sparse Coding based Network (SCN) based on the Learned Iterative Shrinkage and Thresholding Algorithm (LISTA) [16]. Notable improvements are achieved over the generic CNN model in terms of both recovery accuracy and human perception. The implementation is based on recurrent layers that merge linear adjacent ones, allowing to jointly optimize all the layer parameters from end to end. It is achieved by rewriting the activation function of the LISTA layers as follows:

$$[h_\theta(a)]_i = \text{sign}(a_i)\theta_i(\|a_i\|/\theta_i 1)_+ = \theta_i h_1(a_i/\theta_i)$$

Fig. 2 shows the visual difference of a randomly chosen FoodCAT image compared to its SR version. In this example, the original image is 402x125, so the SR was applied with a factor of 3 to assure that both dimensions are bigger than 256.

RESULTS

In this section, we describe the datasets, metrics used for evaluating and comparing each model, and results for each of the image recognition tasks: dishes and food categories.



FIGURE 2. Left shows the SR decreased to 256x256 and right shows the original increased to 256x256.

Dataset

Our dataset, *FoodCAT* has two different labels for each image: Catalan dish, and Catalan food category. Although the total number of Catalan dishes of our datasets are 140, we selected only the set of classes with at least 100 images for our experiments, resulting in a total of 115 classes. Some examples of the available dishes are: sautéed beans, paella, strawberries with vinegar, cuttlefish with peas, roasted snails or beans with sausage. In addition, the images are also labeled in 12 general food categories. Table 1 shows a summary of the general statistics of the dataset, including the number of dishes and images that we have tagged for each food category.

TABLE 1. First column lists the categories, second and third column show the number and the percentage of dishes, and the fourth one shows the amount of pictures by category.

	# dishes	%	# images
Desserts and sweets	34	24,28	11.933
Meats	26	18,57	7.373
Seafood	25	17,85	5.977
Pasta, rice and other cereals	11	7,85	4.728
Vegetables	11	7,85	3.007
Salads and cold dishes	5	3,57	2.933
Soups, broths and creams	8	5,71	2.857
Sauces	4	2,85	2.462
Legumes	6	4,28	1.920
Eggs	5	3,57	615
Snails	3	2,14	470
Mushrooms	2	1,42	438
Total	140	100	44.713

Implementation

There are several frameworks with high capabilities for working on the field of Deep Learning such as TensorFlow, Torch, Theano, Caffe, Neon, etc. We choose Caffe, because it tracks the state-of-the-art in both code and models and is fast for developing. We also decided to use it, because it has a large community giving support on the Caffe-users group and Github, uploading new pre-trained models that people can use for different purposes.

A competitive alternative of the *GoogleNet* model is the *VGG-19*, which we also use in our experiments. This net has 5 blocks of different depth convolutions (64, 128, 256, 512, and 512 consecutively) and 3 FC layers. The first 2 blocks contain 2 different convolutions each and the last 5 contain 4 different convolutions each. It has a total of $2 \times 2 + 3 \times 4 + 3 = 19$ layers. All convolutions have a kernel size of 3×3 with a padding of 1 pixel, i.e. the spatial resolution is preserved after each convolution. Finally, after each convolutional block a max pooling is performed over a 2×2 pixel window with stride 2, i.e. reducing by a factor of 2 the spatial size after each block. As the VGG-19 paper [4] shows, small-size convolution filters are the key to outperform the *GoogleNet* in ILSVRC14 [3] in terms of the single-network classification accuracy.

Evaluation Metrics

Many metrics can be considered to measure the performance of a classification task. In the literature, mainly three methods are used: Accuracy Top-1 (AT1), Accuracy Top-5 (AT5), and the Confusion Matrix (CM). In real-world applications, usually the dataset contains unbalanced classes and the above measures can hide the misclassification of classes with fewer samples. Hence, we consider the Normalized Accuracy Top-1 (NAT1), that gives us the information of how good the classifier is no matter how many samples each class has. Let us define formally each metric.

Let N be the total number of classes with images to test, let N_i be the number of images of the i -th class, and set $n = \sum_{i=0}^{N-1} N_i$, as the total number of images to test. Let $\hat{y}_{i,j}^k$ be the top- k predicted classes of the j -th image of the i -th class, and $y_{i,j}$ the corresponding true class. Let us also define $\mathbf{1}_A : X \rightarrow \{0, 1\}$ as the indicator function as follows:

$$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x_i \in A, \text{ for some } i, \\ 0 & \text{if } x_i \notin A, \text{ for all } i. \end{cases}$$

Then, the definitions of the metrics are as follows:

$$\text{AT1} = \frac{1}{n} \sum_{i,j} \mathbf{1}_{y_{i,j}}(\hat{y}_{i,j}^1), \quad \text{AT5} = \frac{1}{n} \sum_{i,j} \mathbf{1}_{y_{i,j}}(\hat{y}_{i,j}^5), \quad \text{NAT1} = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{N_i} \sum_{j=0}^{N_i-1} \mathbf{1}_{y_{i,j}}(\hat{y}_{i,j}^1).$$

Super Resolution application

For all *FoodCAT* images, we applied the SR method in order to make both image dimensions, width and height, bigger or equal to 256. In Fig. 3, we show the behaviour of the SR algorithm applied on a *Food-101* image. On the left, we show the original image (512×512) resized to the network's input 256×256 , and on the right, we show the same image after resizing it to a smaller resolution than the network's input and applying the SR method for also obtaining a results of 256×256 . Thus, we simulate the result of the SR procedure on *FoodCAT* images: first, improvement through SR and second, resizing to the network's input. We can see that, from a human perception perspective, applying the SR to a low resolution image does not affect the result. Also, when computing the histogram of both images (see Fig. 4), one can see that the difference between them is negligible.

Experimental Results

We need to test the performance of the convolutional neural network on both: dish and food category recognition. **Dish recognition:** One of the richest public domain datasets is the *Food-101* dataset. Since there is small intersection of both datasets, we decided to combine the *FoodCAT* and the *Food-101* dataset in order to build a joint classification model for several types of food. However, in this case we must deal with the differences in image resolution. In order to tackle this problem, we compared the classification on three different dataset configurations (see Fig. 5).

a) *Food-101+FoodCAT*: in this experiment, we use the original images. While all pictures in *Food-101* dataset have similar dimension (width or height) equal to 512, the pictures in *FoodCAT* have a huge diversity in resolutions and do not follow any pattern. On average, their resolution is below 256×256 .

b) *Food-101 halved+FoodCAT*: in this experiment, we decreased the resolution of all images in *Food-101* to make them more alike *FoodCAT*.

c) *Food-101+FoodCAT with SR*: in this experiment, we increased the resolution of all images in *FoodCAT* with the SR technique. Therefore, augmenting the resolution allows to reach a higher fidelity than increasing it with a standard resizing method.



FIGURE 3. Example of SR used in a high resolution image. Left: original image 512x512 resized to 256x256. Right: original image reduced at 40% 230x230, then increased by the SR two times to 460x460, and finally resized to 256x256.

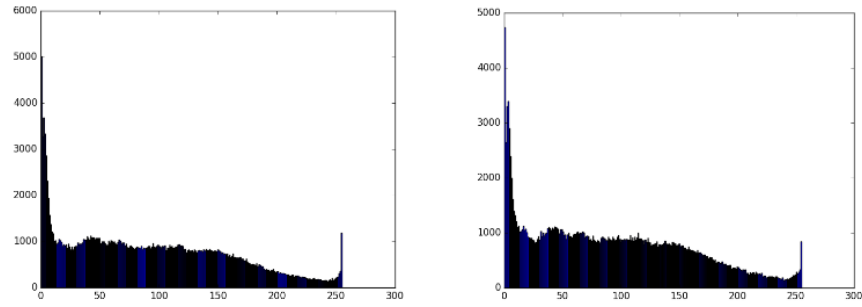


FIGURE 4. Histograms of the original image (left), and the SR (right).

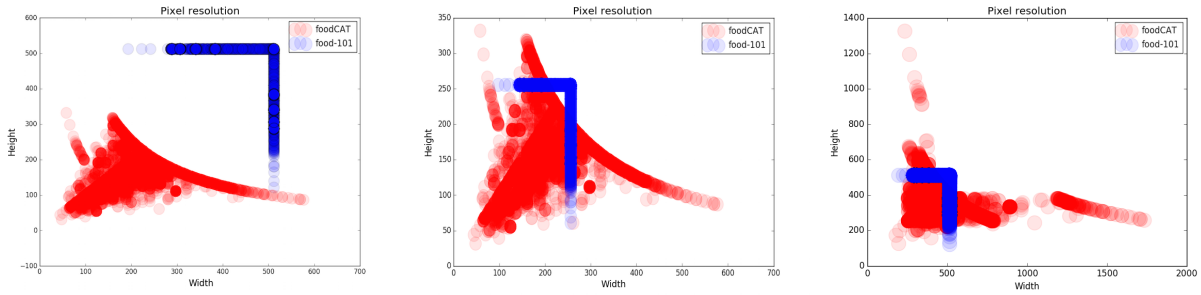


FIGURE 5. Plots of image dimension distributions: left: *Food-101*+*FoodCAT*; center: *Food-101* halved+*FoodCAT* with resolution halved, and right: *Food-101*+*FoodCAT* with SR.

Another of the problems, we have to deal with, when joining two different datasets is the unbalance of classes. Table 2 shows the number of images per learning phase either when using all images (top row) or a maximum of 500 images per class for balance (bottom row).

As a result, dish recognition is performed over *FoodCAT* and *Food-101*, having 115+101 classes to classify respectively. We study the network performance depending on image resolutions and balanced/unbalanced classes. The 6 different experiments are listed below, denoting *GoogLeNet* as 'G' and *VGG-19* 'V':

1. G: *Food-101* + *FoodCAT* with SR.
2. G: *Food-101* + *FoodCAT* with SR, all balanced.
3. G: *Food-101* halved + *FoodCAT*.

TABLE 2. Number of images per learning phase (training, validation and testing) over the complete dataset and the balanced one. The values are presented giving the total number of images in addition to the relative contribution of each dataset in brackets (*Food-101+FoodCAT*).

	training	validation	testing	total
Complete	116.248 (80.800+35.448)	14.540 (10.100+4.440)	14.516 (10.100+4.416)	145.304 (101.000+44.304)
Balanced	73.085 (40.400+32.685)	9.143 (5.050+4.093)	9.124 (5.050+4.074)	91.352 (50.500+40.852)

4. G: *Food-101* halved + *FoodCAT*, all balanced.
5. V: *Food-101* + *FoodCAT*.
6. V: *Food-101* + *FoodCAT*, all balanced.

For all the experiments, we fine-tune our networks after pre-training them on the ImageNet dataset.

Table 3 organises the results of all the 6 different experiments applied either on both datasets ('A, B') or on *FoodCAT* only ('B'). We set the best *AT1*, *AT5*, and *NAT1* in bold, for each of the tested datasets (*Food-101+FoodCAT* or *FoodCAT*). We can see that the best results for the dataset *FoodCAT* (columns 'B') are achieved by a CNN trained from the original dataset (without SR) with balanced classes (experiment 6). It shows the importance of the balanced classes to recognize, with similar accuracy, different datasets with a single CNN. Furthermore, the results of the test in both datasets together (columns 'A, B') are better, when we use all samples in both datasets during the training phase with the method SR applied for the *FoodCAT*. This CNN is the one used in experiment 1, and it also achieves the second best result for the *AT1* over the *FoodCAT* dataset, with a score of 50.02, just 0.57 less than the balanced datasets with VGG (experiment 6). Moreover, adding all scores for the accuracy *AT1* and *AT5*, over the two tests 'A, B' and 'B', experiment 1 has the highest value of 289.44 followed by experiment 6 with value 288.09.

With all this data, we conclude that the best model is the *GoogLeNet* trained from all samples of both datasets, with the SR method applied for *FoodCAT*, corresponding to experiment 1.

TABLE 3. Results of the experiments from 1 to 6. A=*Food-101*, B=*FoodCAT*.

Experiment	1		2		3		4		5		6	
Datasets	A, B	B	A, B	B	A, B	B	A, B	B	A, B	B	A, B	B
<i>AT1</i>	68.07	50.02	62.41	48.94	67.16	49.66	61.28	48.85	67.74	48.12	65.16	50.59
<i>AT5</i>	89.53	81.82	86.81	81.63	89.27	82.07	86.52	80.92	89.28	81.03	88.94	83.40
<i>NAT1</i>	59.08	44.25	57.91	44.44	58.57	44.31	56.99	44.44	58.18	42.34	60.74	46.53

Food categories recognition: The recognition of food categories is performed over the *FoodCAT* dataset by fine-tuning the *GoogLeNet* CNN trained previously with the large dataset ImageNet. We study the network performance depending on if we train all layers or only the last one, the fully-connected layer. Table 4 shows the results obtained for this task. First, if we have a limited machine or limited time, we show that fine-tuning just the fully-connected layer over a model previously trained on a large dataset as *ImageNet* [17], it can give a good enough performance. Training all layers, we achieve recognition of food categories over Catalan food with *AT1* = 72.29 and *AT5* = 97.07. Taking care of the difference of samples on each class, the normalized measure also gives a high performance, with *NAT1* = 65.06.

TABLE 4. Performance and learning time, fine-tuning the *GoogLeNet* model over the food categories labels. We show the results for two experiments done: training all layers, and only training the last fully-connected.

	AT1	AT5	NAT1	# Iterations	Best iteration	Time executing
FC	61.36	93.39	50.78	1.000.000	64.728	12h
All layers	72.29	97.07	65.06	900.000	49.104	24h

Figure 6 shows the normalized Confusion Matrix for the *GoogLeNet* model trained over all layers. It is not surprising that 'Desserts and sweets' is the category that the net can recognize better, as it is also the class with more samples in the dataset with 11.933 images, followed by 'Meats' with 7.373. We also must note that the classes with

less samples in our dataset are 'Snails' and 'Mushrooms', but those specific classes can also be found in the *ImageNet* (the dataset used for the pre-trained model that we are using) that explains the good performance of the network on them.

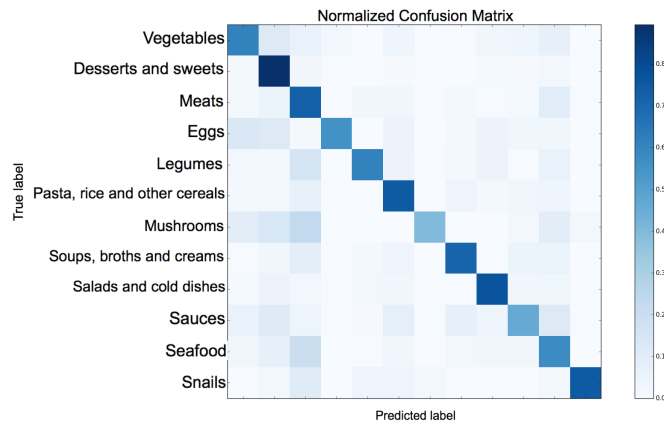


FIGURE 6. Normalized CM of *GoogleNet* model trained over the all layers to recognize food categories.

CONCLUSIONS

In this paper, we presented the novel and challenging multi-labeled dataset related to the Catalan diet called *FoodCAT*. For the first kind of labels, the dataset is divided into 115 food classes with an average of 400 images per dish. For the second kind of labels, the dataset is divided into 12 food categories with an average of 3800 images per dish.

We explored the food classes recognition and found that the best model is obtained by fine-tuning the *GoogleNet* network on the datasets *FoodCAT*, after increasing the resolution with the Super-Resolution method and *Food-101*. This model achieves the highest accuracy top-1 with 68.07%, and top-5 with 89.53%, testing both datasets together, and top-1 with 50.02%, and top-5 with 81.82%, testing only *FoodCAT*. Regarding the food categories recognition, we achieved the highest accuracy top-1 with 72.29% and top-5 with 97.07%, after fine-tuning the *GoogleNet* model for all layers. Our next steps are to increase the dataset and explore other architectures of convolutional neural networks for food recognition.

ACKNOWLEDGMENTS

This work was partially funded by TIN2015-66951-C2-1-R, La Marató de TV3, project 598/U/2014 and SGR 1219. P. Radeva is supported by an *ICREA Academia* grant. Thanks to the University of Groningen for letting us use the Peregrine HPC cluster.

REFERENCES

- [1] F. Monteiro-Silva. Olive oil's polyphenolic metabolites - from their influence on human health to their chemical synthesis. *ArXiv e-prints 1401.2413*, January 2014.
- [2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL <http://arxiv.org/abs/1409.4842>.
- [3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

- [4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, arXiv:1409.1556, 2014.
- [5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- [6] Y. Kawano and K. Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.
- [7] Atsushi Tatsuya and Aono Masaki. Food image recognition using covariance of convolutional layer feature maps. *IEICE TRANSACTIONS on Information and Systems*, 99(6):1711–1715, 2016.
- [8] Marc Bolaños and Petia Radeva. Simultaneous food localization and recognition. In *Proceedings of the International Conference on Pattern Recognition (in press)*, 2016. URL <http://arxiv.org/abs/1604.07953>.
- [9] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo, ICME 2012, Melbourne, Australia, July 9-13, 2012*, pages 25–30, 2012. doi: 10.1109/ICME.2012.157. URL <http://dx.doi.org/10.1109/ICME.2012.157>.
- [10] Institut Català de la Cuina. *Corpus del patrimoni culinari català*. Edicions de la Magrana, 2011. ISBN 9788482649498.
- [11] Hajime Hoashi, Taichi Joutou, and Keiji Yanai. Image recognition of 85 food categories by feature fusion. In *12th IEEE International Symposium on Multimedia, ISM 2010, Taichung, Taiwan, December 13-15, 2010*, pages 296–301, 2010. doi: 10.1109/ISM.2010.51. URL <http://dx.doi.org/10.1109/ISM.2010.51>.
- [12] Taichi Joutou and Keiji Yanai. A food image recognition system with multiple kernel learning. In *Proceedings of the 16th IEEE International Conference on Image Processing, ICIP'09*, pages 285–288, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-5653-6. URL <http://dl.acm.org/citation.cfm?id=1818719.1818816>.
- [13] Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. Deep image: Scaling up image recognition. *CoRR*, arXiv:1501.02876, 2015. URL <http://arxiv.org/abs/1501.02876>.
- [14] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 370–378, 2015.
- [15] X. Jin, Y. Chen, J. Dong, J. Feng, and S. Yan. Collaborative Layer-wise Discriminative Learning in Deep Neural Networks. *ArXiv e-prints*, July 2016.
- [16] Johannes Fürnkranz and Thorsten Joachims, editors. *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, 2010. Omnipress.
- [17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.